Applications of item response theory to non-cognitive data

Egberink, Iris Jo-Anne Lieneke

Chapter 2

An Item Response Theory Analysis of Harter's Self-Perception Profile for Children or Why Strong Clinical Scales Should Be Distrusted

Abstract

We investigated the psychometric properties of the subscales of the Self-Perception Profile for Children with item response theory (IRT) models using a sample of 611 children. Results from a nonparametric Mokken analysis and a parametric IRT approach for boys (n = 268) and girls (n = 343) were compared. We found that most scales formed weak scales and that measurement precision was relatively low and only present for latent trait values indicating low self-perception. The subscales Physical Appearance and Global Self-Worth formed one strong scale. Children seem to interpret Global Self-Worth items as if they measure Physical Appearance. Furthermore, we found that strong Mokken scales (such as Global Self-Worth) consisted mostly of items that repeat the same item content. We conclude that researchers should be very careful in interpreting the total scores on the different Self-Perception Profile for Children scales. Finally, implications for further research are discussed.

This chapter has been accepted for publication as:
Egberink, I. J. L., & Meijer, R. R. (in press). An IRT analysis of Harter's Self-Perception Profile for Children or why strong clinical scales should be distrusted. *Assessment.*

2.1 Introduction

A homogeneous, Guttman-type test yields high precision, but covers little ground. (…). It is worth sacrificing fidelity to attain (...) bandwith (Cronbach, 1954, pp. 268-270)

Self-perception or self-concept is an important construct in developmental psychology, although there is still a debate about the exact relation between self-perception and important outcome variables. Self-perception is often used to refer to the way people evaluate their various abilities and attributes (e.g., Harter, 1985). Some authors claim that positive feelings about the self are important for healthy developmental outcomes, such as subjective well-being (e.g., DeNeve & Cooper, 1998; Diener & Diener, 1995), healthy social relationships and attachment (e.g., Leary, Tambor, Terdal, & Downs, 1995; Murray, Holmes, & Griffin, 2000), and academic achievement and occupational success (e.g., Elliott, 1996; Hansford & Hattie, 1982; Judge & Bono, 2001). Negative feelings of the self are claimed to be related to problematic outcomes, such as poorer mental and physical health (e.g., Roberts, Gotlib, & Kassel, 1996; Trzesniewski et al., 2006) and antisocial behavior (e.g., Donnellan, Trzesniewski, Robins, Moffitt, & Caspi, 2005; Rosenberg, Schooler, & Schoenbach, 1989). Recent evidence for the importance of self-concept is presented in, for example, Montague, Enders, Dietz, Dixon, and Morrison Cavendish (2008), who found a strong relation between depressive symptoms and self-perception in a school-based sample at risk for developing emotional and behavioral disorders. They found that when depressive symptoms decreased self-perception improved. Also, a low self-concept may be related to the development of obesity (e.g., Schumann et al., 1999).

Baumeister, Campbell, Krueger, and Vohs (2003) reviewed the self-esteem literature and concluded that self-esteem decreases the chances to become depressed and to develop eating disorders and is positively associated with life satisfaction. However, they found that it is unclear whether self-esteem determines positive social relations, academic achievement, and academic success.

Dubois and Tevendale (1999; see also Kernis, 2006) presented a critical review of the self-concept and self-esteem debate, where two opposite views were discussed, namely self-esteem as a "vaccine", that is, as a construct that prevents against educational failure and crime, or as a "epiphenomenon", that is, as a construct that indicates that "the interaction with the world is not going well". Dubois and Tevendale (1999) concluded that self-concept could be investigated best when considered consisting of different distinct concepts and that the moderating

influence of youth characteristics and environmental processes should be taken into account.

Harter's (1985) Self-Perception Profile for Children (SPPC; Veerman, Straathof, Treffers, van den Bergh, & ten Brink, 2004) is a popular instrument to measure self-concept. It is often used as a measure to determine self-concept in schools and clinical treatment centers. This self-report inventory is intended to determine how children between 8 and 12 years of age judge their own functioning in several specific domains and how they judge their global self-worth. The SPPC consists of six subscales, each consisting of six items. Five of the six subscales represent specific domains of self-concept: Scholastic Competence (SC), Social Acceptance (SA), Athletic Competence (AC), Physical Appearance (PA), and Behavioral Conduct (BC). The sixth scale measures Global Self-Worth (GS), which according to Harter (1985) is a more general concept. When a child fills out the SPPC, he or she first chooses which of the two statements applies to him or her and then indicates if the chosen statement is "sort of true for me" or "really true for me". Scoring is done on a 4-point scale. The answer most indicative of a positive self-concept is scored "4", and the answer least indicative of competence is scored "1".

The psychometric properties of the five specific domain scales of the SPPC have been investigated using classical test theory and factor analytical approaches. Both research on the original English version and research on translated versions showed that a 5-factor model gave a reasonable fit. For example, Granleese and Joseph (1993) replicated the 5-factor structure obtained previously by Harter (1985) with American adolescents. Furthermore, they found a strong similarity in the correlations between subscale scores for girls and boys. Schumann et al. (1999) evaluated the SPPC in a biracial cohort for third graders and found a 5-factor solution for White girls, but not for Black girls. For Black girls, the physical appearance and the athletic competence domains were not yet fully differentiated. Thill et al. (2003) compared the SPPC with a questionnaire on actual behavior for children with and without spina bifida, and they found that the factor structure was similar for both groups.

For the Dutch version, Veerman et al. (2004) found a reasonable fit of the 5-factor model, where coefficient alpha for the subscales ranged from .68 (BC) to .83 (PA). In their study, van den Bergh and van Ranst (1998) also analyzed the Dutch version of the SPPC. They found that the factorial structure of the underlying self-concept was not exactly the same for fourth and sixth graders and that the SPPC was less reliable for boys than for girls and suggested that when performance of a

specific child has to be evaluated, the child is best situated in his or her gender group.

Recently, Meijer, Egberink, Emons, and Sijtsma (2008) used the SPPC to illustrate the usefulness of studying individual item score patterns. They identified children for whom the SPPC did not reflect their self-concept as a result of cognitive deficits. However, both the Meijer et al. (2008) study and the van den Bergh and van Ranst study (1998) suggested that the psychometric quality of the individual scales differed and that a more thorough psychometric analysis was needed to obtain a better picture of the characteristics of these scales.

The aim of the present study was to assess the psychometric quality of the SPPC scales by means of item response theory (IRT; Embretson & Reise, 2000) models. By using IRT modeling, a more detailed picture can be obtained about the functioning of individual items and scales and about the relation between trait scores (in this study self-concept scores) and item endorsement. Classical test theory and factor analytical approaches do not provide exhaustive item-level analysis. In the present study, we were in particular interested in (a) which items in each scale mainly determine the construct that is being measured and (b) whether the scales can reliably distinguish persons across different values of the latent trait scale. With IRT it is possible to investigate the relative contribution of each item to the measurement precision of the scale and to determine which items are most related to the construct being investigated. Thus, IRT can be used to obtain more refined information about the construct validity of the scales.

2.2 Method

2.2.1 Participants and Procedure

Part of the data that were reported in Meijer et al. (2008) were reanalyzed. Data were collected from 611 children between 8 and 12 years of age. The sample contained 343 girls (mean age = 10.19, *SD* = 1.29) and 268 boys (mean age = 10.17, *SD* = 1.23). Most children were White. These children came from five primary schools in the east of the Netherlands. Two schools are public primary schools and three are Catholic primary schools. Of the 611 children, there were 45 second graders (mean age = 8.33, *SD* = 0.37), 126 third graders (mean age = 8.73, *SD* = 0.47), 139 fourth graders (mean age = 9.79, *SD* = 0.58), 155 fifth graders (mean age = 10.73, *SD* = 0.50), and 146 sixth graders (mean age = 11.80, *SD* = 0.55). The research reported in

this study was part of a larger project in which information was obtained routinely from the children about their emotional and personal well-being.

2.2.2 Item Response Theory

IRT models are based on the idea that psychological constructs are latent, that is, not directly observable, and that knowledge about these constructs can only be obtained through the manifest responses of persons to a set of items (e.g., Embretson & Reise, 2000; Sijtsma & Molenaar, 2002). IRT explains the structure in the manifest responses by assuming the existence of a latent trait, denoted by the Greek letter θ. By means of IRT models, it is possible to locate a person's θ and the characteristics of the items that make up the measurement instrument on the same metric (i.e., latent trait continuum).

In IRT, both nonparametric and parametric approaches can be distinguished. Nonparametric IRT models are based on less restrictive assumptions about the data and are, therefore, ideal instruments to explore the psychometric structure of tests and questionnaires. Parametric approaches are based on more restrictive assumptions but provide information that cannot be obtained using nonparametric approaches. In this study, we used Mokken's nonparametric monotone homogeneity model (MMH; Sijtsma & Molenaar, 2002) to explore the psychometric structure of the SPPC and the parametric graded response model (GRM; Samejima, 1969, 1997) to obtain more detailed information about the measurement precision of the SPPC scales across the latent trait continuum. Furthermore, we used both approaches to obtain a detailed picture about the psychometric quality of the scales.

Mokken Scaling

The MMH is based on the assumptions of unidimensionality, local independence, and monotonicity. The model assumes that all items in a test or questionnaire measure the same latent trait (unidimensionality assumption), that a person's response to one item is not influenced by the response to another item (local independence), and that the item response function is nondecreasing (monotonicity assumption). A more detailed description of these assumptions can be found in Sijtsma and Molenaar (2002) or Meijer and Baneke (2004).

To check the assumptions of the MMH, several methods have been proposed. In this study, we used the coefficient H_i for items and the coefficient H for a set of items. Under the MMH, higher positive H values reflect higher discrimination power of the items, and as a result, more confidence in the ordering of respondents by means of their total scores. This is also referred to as scalability, that is, the degree to

which a set of items are related to each other and form a scale. Items with high H_i values discriminate well in the group in which they are used. H_i values determine how well an item fits the scale. For practical test construction purposes, the following rules of thumb have been suggested. Weak scalability is obtained if $.3 \leq H < .4$, medium scalability if $.4 \leq H < .5$, and strong scalability if $.5 \leq H < 1$ (Sijtsma & Molenaar, 2002). Values of H smaller than .3 are considered evidence that the items are unscalable for practical purposes.

We used the computer program Mokken Scale Analysis for Polytomous Items version 5.0 for Windows (MSP5.0; Molenaar & Sijtsma, 2000) to conduct a Mokken scale analysis for each subscale of the SPPC. We checked the assumptions of the MMH by inspecting the H and H_i coefficients. Because we suspected on the basis of the literature (van den Bergh & van Ranst, 1998) and on the basis of our own observations during test administration that there may be differences in model fit for boys and girls, we ran separate analyses for boys and girls.

Graded Response Model

To obtain a more detailed picture of the psychometric quality of the SPPC, we also analyzed the data with the GRM (Samejima, 1969, 1997). The GRM is suitable for analyzing ordered response categories, such as Likert-type rating scales. Several researchers used this model to analyze personality data, and there is a close relationship between the GRM and Mokken's MMH model (Sijtsma & Molenaar, 2002). The MMH can be interpreted as a nonparametric version of the GRM, in the sense that both models assume unidimensional measurement, local independence, and nondecreasing item response functions.

The items in the GRM are defined by a discrimination parameter (α; usually with numerical values between 0.5 and 2.5) and two or more location parameters (β_m; usually with numerical values between -2.5 and +2.5); the number of location parameters per item is equal to the number of response categories minus 1; thus, in our analysis, $4 - 1 = 3$. Like the H_i coefficient, the magnitude of the discrimination parameter reflects the degree to which the item is related to the underlying latent trait. This means that for high α values the response categories accurately differentiate among trait levels. The location parameters reflect the spacing of the ordered response categories along the θ scale. The location parameter β_m can be interpreted as the point at the latent trait continuum where there is a 50% chance of scoring in category m or higher. Thus, respondents with a θ value higher than β_m have more than 50% chance of responding in category m or higher. These α and β_m parameters are used to determine the probability of an examinee to respond in a

particular response category. These probabilities can be used to determine the category response functions, which describe the probability of responding in a particular response category conditional on θ.

Figure 2.1 gives an example of the category response functions for two items of the BC scale of the SPPC for girls, Item 6 with a high estimated α value ($\hat{\alpha} = 2.38$; upper panel) and Item 1 with a low estimated α value ($\hat{\alpha} = 0.76$; lower panel). Moving from the lower to the higher end of the θ scale shows that first Category 1 is most likely (low θ levels), then Category 2, followed by Category 3, and, finally, Category 4 (high θ level). Furthermore, the middle category options are more peaked (higher $\hat{\alpha}$ values) for Item 6 than for Item 1.

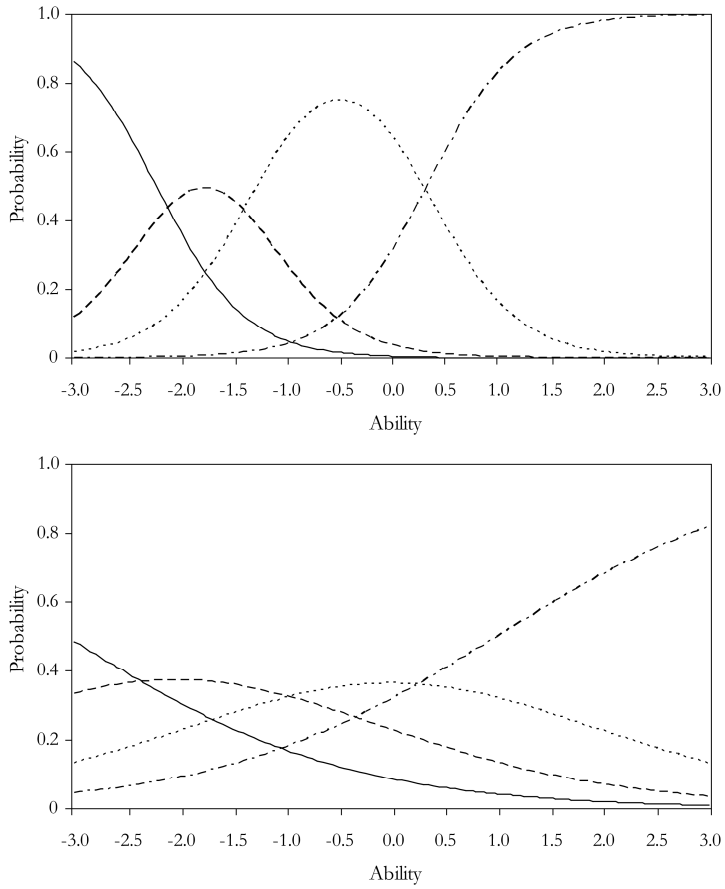

Figure 2.1: Category response functions for Item 6 of the Behavioral Conduct scale with $\hat{\alpha} = 2.38$ ($H_i = .39$; upper panel) and Item 1 of the Behavioral Conduct scale with $\hat{\alpha} = 0.76$ ($H_i = .22$; lower panel) for girls.

An important difference between Mokken scaling and the GRM is that in the former persons are assumed to have equal standard errors regardless of their position on the construct. In Mokken scaling, like in classical test theory, there is one reliability estimate. In parametric IRT, the concept of reliability is replaced by the concepts of item and test information. The standard error of a trait estimate is inversely related to the square root of the test information function. Thus, persons may have different standard errors depending on how discriminating a set of items is in different ranges of the latent trait. In general, items with larger discrimination parameters (i.e., the α parameters) provide relatively more information. The location parameters (i.e., the β_m parameters) determine where the information is located. Item information is additive across the items administered and test information is maximized around the location parameters. Because information is inversely related to the standard error of measurement, this feature of IRT allows us to determine how precise a measure is for individuals in high, medium, and low trait ranges. We estimated the item parameters for the GRM using MULTILOG 7.0 (Thissen, Chen, & Bock, 2003).

2.3 Results

2.3.1 Descriptive Statistics and Nonparametric Scaling

Table 2.1 depicts the mean item scores, item-test correlation, coefficient alpha, Guttman's lambda-2, H, and H_i coefficients for girls and boys[1]. Like coefficient alpha, Guttman's lambda-2 is a lower bound to test reliability, but it can be shown that Guttman's lambda-2 is equal or greater than coefficient alpha and a more accurate lower bound than coefficient alpha. A first observation is that mean item scores are high, that is, most children have a positive self-concept on the respective subscales. Furthermore, we observe that there are differences in the psychometric quality of the different scales for both boys and girls and between boys and girls. This is reflected in the different mean values of the item-test correlations and H values for the different scales.

In general, items are less scalable for boys than for girls. For boys, SC, SA, PA and BC form weak scales ($.3 \leq H < .4$, although some items have $H_i < .3$; strictly

[1] There were also some differences between the younger children and older children. For very young children, the items were less scalable than for older children. However, the very young children constitute only a small part of the sample. See Meijer et al. (2008) for more details.

Table 2.1

Descriptive Statistics for Girls and Boys.

		Girls					Boys				
Item		λ_2 [a]/ α [b]	*M*	r_{it}	H_i	*H*	λ_2 [a]/ α [b]	*M*	r_{it}	H_i	*H*
SC	1		2.49	.43	.34			2.80	.47	.36	
	2		2.67	.45	.35			2.92	.51	.38	
	3		2.78	.56	.42			2.75	.56	.42	
	4		2.92	.50	.38			2.93	.48	.36	
	5		3.23	.54	.44			3.27	.53	.42	
	6		2.81	.59	.44			2.94	.60	.44	
		.77/.77				.39	.78/.77				.39
SA	7		3.20	.50	.38			3.36	.49	.36	
	8		3.23	.61	.45			3.42	.61	.44	
	9		2.77	.52	.40			2.87	.52	.38	
	10		3.27	.43	.33			3.33	.37	.28	
	11		2.86	.48	.37			2.88	.46	.34	
	12		3.18	.51	.39			3.18	.35	.27	
		.77/.76				.39	.74/.73				.35
AC	13		2.98	.43	.30			3.07	.25	.17	
	14		2.83	.47	.33			2.93	.26	.18	
	15		3.10	.38	.27			3.25	.37	.24	
	16		2.62	.52	.38			2.90	.39	.26	
	17		3.41	.28	.22			3.42	.34	.23	
	18		2.99	.42	.31			3.00	.38	.25	
		.69/.68				.30	.59/.59				.22
PA	19		3.26	.60	.50			3.49	.50	.37	
	20		3.05	.54	.46			3.30	.35	.27	
	21		3.22	.71	.57			3.35	.56	.40	
	22		3.17	.74	.59			3.37	.60	.43	
	23		3.21	.58	.48			3.42	.52	.38	
	24		3.30	.70	.58			3.46	.58	.43	
		.86/.85				.53	.77/.77				.38
BC	25		2.87	.29	.22			2.85	.35	.26	
	26		3.28	.49	.36			3.10	.51	.36	
	27		3.19	.40	.29			3.16	.32	.25	
	28		2.93	.41	.30			2.85	.43	.32	
	29		2.93	.48	.35			2.70	.55	.40	
	30		3.21	.56	.39			3.10	.58	.41	
		.70/.70				.31	.72/.72				.33
GS	31		3.23	.59	.48			3.43	.27	.20	
	32		3.35	.51	.41			3.40	.42	.30	
	33		3.50	.60	.47			3.65	.43	.31	
	34		3.47	.64	.50			3.54	.54	.37	
	35		3.39	.65	.50			3.56	.48	.33	
	36		3.14	.44	.37			3.17	.31	.25	
		.81/.81				.45	.69/.67				.29

Note. *M* = mean item score; r_{it} = item-test correlation; SC = Social Competence; SA = Social Acceptance; AC = Athletic Competence; PA = Physical Appearance; BC = Behavioral Conduct; GS = Global Self-Worth. [a] = Guttman's lambda-2. [b] = coefficient alpha.

speaking only the SC scale forms a weak scale). The worst scale is the AC scale with H_i values between .17 and .26; thus boys are unscalable with respect to AC. For girls, in general, scalability coefficients are higher than for boys; for PA and GS they are much higher. These scales can be characterized as medium and strong scales. Intercorrelations between the total scores on the five basic scales ranged from r = .17 through r =.43 for boys and from r =.21 through r = .41 for girls. Correlations between the basic scales and GS ranged, however, from .24 (AC) to .68 (PA) for boys and between .33 (SC) and .76 (PA) for girls. At the end of the results section, we discuss IRT results when we combine items of the PA and GS scales.

2.3.2 Parametric IRT Analysis

In Table 2.2, we depicted the estimated item parameters for girls and boys. There are some interesting observations. A first observation is that there are items with extremely high $\hat{\alpha}$ values. For example, consider the PA scale for girls, here two items, Item 21 and Item 24, have $\hat{\alpha}$ values near 3.0, and a third item, Item 22, has an $\hat{\alpha}$ value of 3.55. One possibility is that this may point at item content redundancy, which is asking the same question twice. Items 21 and 22 use exactly the same phrasing but differ in one word: the Dutch word "lichaam" (English translation "body") and the Dutch word "uiterlijk" (English translation "appearance"). So there is a subtle difference between the two items. Item 24 seems to be the item that is the shortest summary of PA ("I look good").

Another possibility as suggested by Reise and Waller (2009) is that there is a highly skewed construct or quasi-trait. Many constructs in psychology, although assumed dimensional, are possibly, what they called "quasi-traits", that is, traits that are only defined at one end of the latent trait scale. Reise and Waller mentioned constructs such as self-esteem, aggression, and spirituality. Self-perception as measured by the SPPC, which is conceptually related to self-esteem, also seems to be a quasi-trait. Consider the location parameters given in Table 2.2. For all scales the $\hat{\beta}_2$ parameters are negative. Remember that persons with an estimated theta value, denoted by $\hat{\theta}$, higher than $\hat{\beta}_2$ have more than 50% chance of responding in Category 2 or higher. Thus, persons with $\hat{\theta}$ = 0, that is, with a mean score on self-concept, have more than 50% chance to answer in Category 3, and for the PA and the GS scale even in Category 4. This clearly indicates that the item locations are at the left side of the latent trait scale. One explanation may be due to the nature of the self-perception construct; items only differentiate between children with low self-perception because researchers are mainly interested in this end of the construct. A

Table 2.2

Estimated Item Parameters for Girls and Boys.

		Girls				Boys			
Item		$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$
SC	1	1.15	-1.78	0.01	1.70	1.28	-2.00	-0.56	0.88
	2	1.24	-2.00	-0.16	1.06	1.48	-2.05	-0.64	0.58
	3	1.76	-1.73	-0.49	0.92	1.90	-1.50	-0.36	0.72
	4	1.55	-2.16	-0.70	0.70	1.46	-2.07	-0.72	0.56
	5	1.91	-2.54	-1.32	0.31	1.69	-3.35	-1.60	0.32
	6	1.98	-1.70	-0.54	0.86	2.16	-1.83	-0.61	0.51
SA	7	1.51	-2.15	-1.17	0.01	1.87	-1.99	-1.33	-0.37
	8	2.47	-1.66	-0.97	-0.12	3.28	-1.76	-1.16	-0.36
	9	1.63	-1.32	-0.49	0.65	1.74	-1.34	-0.59	0.38
	10	1.29	-2.92	-1.55	0.07	1.15	-3.04	-1.85	-0.08
	11	1.44	-1.74	-0.72	0.70	1.27	-1.83	-0.74	0.57
	12	1.52	-2.70	-1.39	0.43	0.96	-3.26	-2.04	0.54
AC	13	1.46	-1.80	-0.81	0.24	0.89	-2.20	-1.33	-0.11
	14	1.70	-1.10	-0.46	0.17	0.79	-2.08	-1.00	0.08
	15	1.18	-2.62	-1.30	0.40	1.51	-2.57	-1.39	0.09
	16	2.02	-1.55	-0.22	1.20	1.20	-2.67	-0.97	1.18
	17	0.81	-4.03	-2.36	-0.64	1.30	-2.50	-1.68	-0.68
	18	1.25	-2.55	-1.10	0.79	1.52	-2.16	-0.98	0.54
PA	19	1.99	-2.07	-1.12	-0.02	1.63	-2.52	-1.77	-0.53
	20	1.57	-1.62	-0.67	0.07	0.98	-2.59	-1.48	-0.63
	21	2.81	-1.49	-0.75	-0.21	1.93	-1.86	-1.09	-0.49
	22	3.55	-1.46	-0.65	-0.09	2.73	-1.78	-0.99	-0.42
	23	2.15	-1.61	-0.87	-0.12	2.18	-1.86	-1.22	-0.54
	24	2.99	-1.88	-1.09	0.01	2.31	-2.24	-1.52	-0.32
BC	25	0.76	-3.08	-1.01	1.00	0.95	-2.47	-0.77	0.87
	26	1.97	-2.62	-1.45	0.19	1.74	-2.43	-1.16	0.54
	27	1.25	-3.26	-1.63	0.46	1.00	-3.22	-1.75	0.49
	28	1.09	-2.69	-0.74	0.61	1.16	-2.37	-0.49	0.72
	29	1.73	-2.26	-0.73	0.68	1.79	-1.65	-0.30	0.91
	30	2.38	-2.22	-1.31	0.33	2.58	-2.17	-1.09	0.55
GS	31	1.87	-2.31	-1.19	0.17	0.91	-3.85	-2.71	-0.39
	32	1.49	-2.22	-1.41	-0.37	1.46	-2.12	-1.49	-0.64
	33	2.53	-2.28	-1.42	-0.40	2.03	-2.61	-2.10	-0.79
	34	2.94	-2.06	-1.26	-0.41	3.05	-1.76	-1.39	-0.70
	35	3.19	-1.84	-1.06	-0.32	2.40	-2.08	-1.57	-0.70
	36	1.21	-2.82	-1.25	0.33	0.89	-3.69	-1.86	0.48

Note. SC = Social Competence; SA = Social Acceptance; AC = Athletic Competence; PA = Physical Appearance; BC = Behavioral Conduct; GS = Global Self-Worth.

high level of self-perception (i.e., high self-concept) is healthy, whereas a low level of self-perception (i.e., low self-concept) may be problematic. Therefore, the aim of measuring self-perception is to detect children with low self-perception, and therefore, items are written that measure low self-perception. An alternative interpretation is that most people in today's Western world have high self-esteem (see Baumeister, 1993).

Another interesting observation concerns the SA and the GS scales for boys. For the SA scale, Item 8 ("Having many friends") has a very high discrimination parameter ($\hat{\alpha} > 3$), whereas Items 10, 11, and 12 have low discrimination parameters (around $\hat{\alpha} = 1$). Conceptually this means that Item 8 defines SA, whereas Items 10, 11, and 12 are much less related to SA. For example, Item 10, "Doing many things alone or with others", may indicate social acceptance but may also be related to a preferred way of doing things. It is clear that, in general, high α values go together with high H_i as expected. In this data set, $H_i \geq .30$ corresponds to $\hat{\alpha} \geq 1.20$ and $H_i \geq .40$ correspond to $\hat{\alpha} \geq 1.50$. Although H_i coefficients and α parameters both indicate the slope of the item response function, both statistics are also sensitive to different characteristics of the data. H_i is strongly influenced by the probability distribution of the latent trait values, which also prevents a researcher from using a scale in a population where it cannot discriminate between persons. In the literature there is a strong emphasize on selecting items with H_i values larger than some lower bound as, say, $H_i = .3$. We observe, however, that a researcher should also be careful when H_i values are very high. We agree with Sijtsma and Molenaar (2002) that "one should find measurement instruments that measure one *meaningful* psychological ability or trait at a time" (p. 19). The question here is what is "meaningful". As we discussed, repeating items with a similar content will result in scales with high H values but, sometimes, extremely small-band constructs. It is, therefore, very important to inspect the content of the items and the scales and, perhaps most important, how the content of the items is interpreted. Especially for special groups such as young children or clinical patients, items may be interpreted differently than a researcher is suspecting.

Because large H_i values and large α parameters go together, strong Mokken scales may also be the result of violations of local independence and the result of narrow-band constructs. Although these scales are very reliable, one should be careful to include items in a scale that are not semantically similar. High H_i values may also point at items that define the construct ("I am often depressed" in a depression list). For example, Emons, Meijer, and Denollet (2007) analyzed Negative Affectivity items and found that the dysphoria item "is often down in the dumps" had an $\hat{\alpha}$

parameter value of 3.61, whereas an item such as "is easily irritated" had an $\hat{\alpha}$ parameter value of 1.45.

2.3.3 Measurement Precision

Figure 2.2 displays the test information functions for the six subscales of the SPPC for girls (for boys similar results were obtained). For all scales the highest information is located at the lower trait values, that is, between scale scores $\hat{\theta}$ = -2 and 0. This can be explained by noting that most item locations are situated at the lower trait ranges, that is, all items are relatively easy or popular. Remember that information is inversely related to the standard error of measurement and this feature thus allows us to determine the measurement precision of the SPPC scales for children in low, medium, and high trait ranges.

From Figure 2.2, it is clear that for girls the scales do *not* provide precise measurement across the whole scale and that even in the parts of the scale where there is some measurement precision (with the exception of the PA and the GS scales) broad confidence intervals result. For example, the six items that together measure AC provide information of around 4 for values between $\hat{\theta}$ = -2 and 0, which corresponds to *SE* = 0.50. Thus, for a child with $\hat{\theta}$ = -1, the 95% confidence interval is between -2 and 0, which is clearly too broad to base any substantive conclusion on. Even for the subscales that together form a weak Mokken scale, such as SC and SA, using only six items often result in very broad confidence intervals. For example, for SC and scale score $\hat{\theta}$ = 0 information equals 5, which corresponds to *SE* = .45 and thus a 95% confidence band between -.90 and +.90. Similar results were obtained for boys. Thus, the take-home message is that these subscales only provide some measurement precision at the left part of the θ scale (i.e., for total scores smaller than 15, which corresponds to $\hat{\theta}$ = 0) and that even these score profiles should be interpreted very carefully because for most scales broad confidence bands exist.

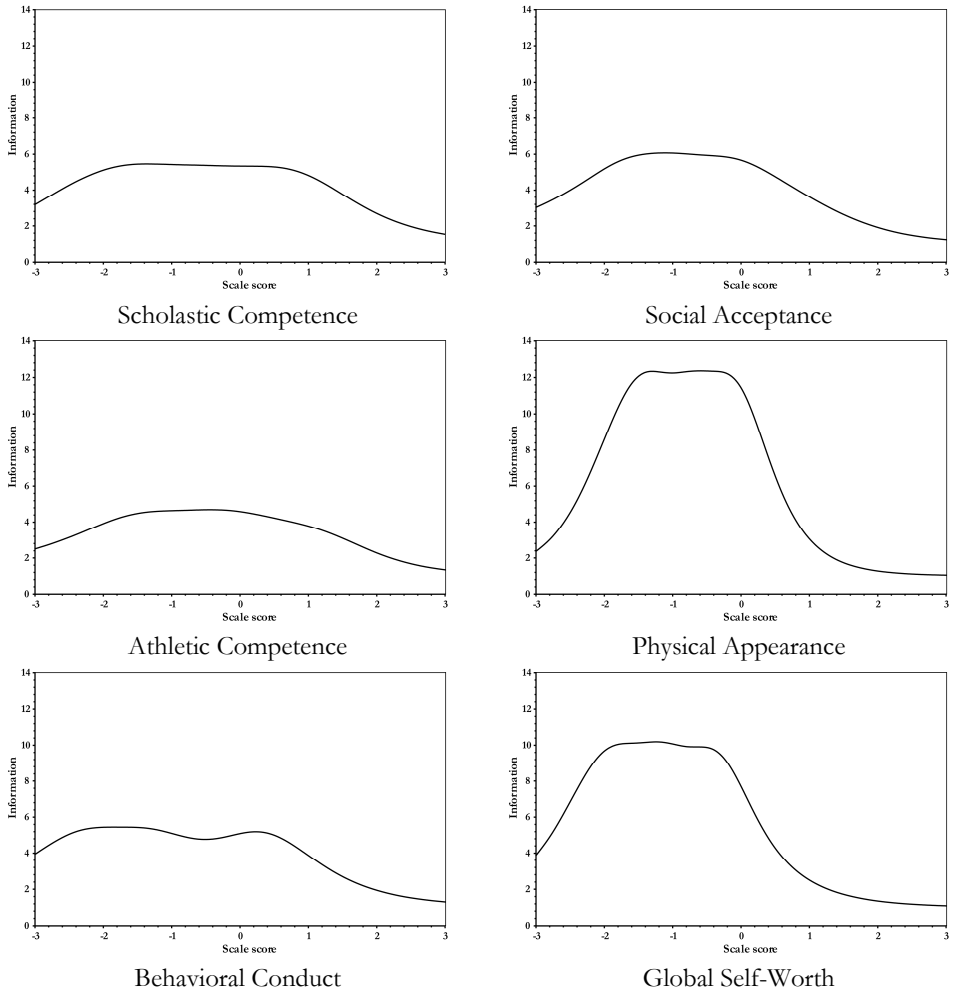

Figure 2.2: Test information functions for the six subscales of the SPPC for girls.

2.3.4 Combined Scales

Because correlations between the PA and GS scales were high, we investigated whether items of these scales cluster together. Both for boys and for girls this was the case. Consider the item parameters and H_i values in Table 2.3 for the combined scale (PA and GS together) for girls. For most items, the H_i values of the combined scale are comparable with the H_i values for the individual scales (depicted in Table 2.1). For 7 out of the 12 items, the $\hat{\alpha}$ parameters for the combined scale were higher than for the individual scales (Table 2.2). This clearly indicates item redundancy, because it is expected that when we measure a broader trait as a result of combining two scales, item discrimination will decrease.

Table 2.3

Estimated Item Parameters and H_i Coefficients for 12 items of the GS and PA Scale as One Scale for Girls (H = .48, When Item 32 Is Removed, H = .50).

Item		$\hat{\alpha}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	H_i
PA	19	2.13	-1.79	-0.89	0.14	.48
PA	20	1.55	-1.43	-0.51	0.22	.41
PA	21	2.60	-1.31	-0.59	-0.05	.51
PA	22	3.65	-1.23	-0.48	0.06	.57
PA	23	2.23	-1.38	-0.69	0.03	.45
PA	24	3.32	-1.60	-0.86	0.16	.55
GS	31	1.76	-2.16	-1.03	0.33	.45
GS	32	1.25	-2.27	-1.36	-0.22	.36
GS	33	2.59	-2.03	-1.19	-0.21	.51
GS	34	2.69	-1.88	-1.07	-0.24	.51
GS	35	4.14	-1.50	-0.81	-0.12	.57
GS	36	1.24	-2.55	-1.04	0.49	.35

Note. PA = Physical Appearance; GS = Global Self-Worth.

Another observation is that half of the amount of information is due to three items. In Figure 2.3, we provide the test information curve for all 12 items in the combined scale. Test information equals 23 between $\hat{\theta}$ = -2 and 0. When we select Item 22 ("Satisfied with how I look like") and Item 24 ("Happy with how I look like") of the PA scale, and Item 35 of the GS scale ("Happy with who I am"), we obtain test information of 12. These three items provide half of the information of a scale that is 4 times as long. Thus, the combination of GS items and PA items is for a large part a "Happy with how I look like" scale.

We further conclude that, although some authors echo Harter's (1985) idea that Global Self-Worth "is not a measure of general competence and as such should be considered an independent component of the scale" (e.g., Shevlin, Adamson, & Collins, 2003, p. 1995), we found high correlations between PA and GS. This was also found by Schumann et al. (1999) for girls. GS is thus not an independent component of the scale, but is heavily saturated with PA. For children, physical appearance heavily determines global self-worth. This close association between satisfaction with physical appearance and global self-worth has been hypothesized to contribute to the risk for the development of eating disorders in adolescent girls and may result from an overemphasis by the popular media on the importance of looking good.

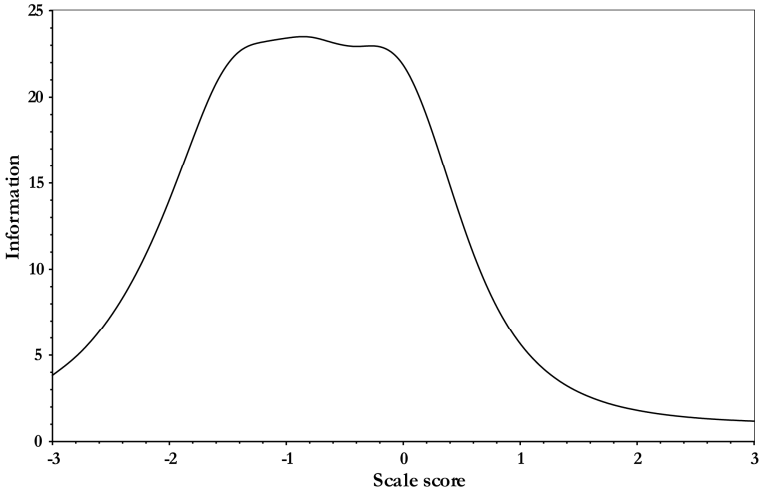

Figure 2.3: Test information function for the combined scale (Physical Appearance and Global Self-Worth scales together) for girls.

2.4 Discussion

What did we learn from our analyses? Our analyses suggest that when interpreting SPPC scores, care should be taken in interpreting the different subscale scores. First, measurement precision differs across scales and across the latent trait values within a scale. Like in many typical performance measures (see Reise & Waller, 2009), self-perception is only defined at one end of the scale, that is, self-perception in a non-clinical population is not a bipolair trait. The trait is only defined at the lower end of the trait scale, where children are situated with a relatively low self-concept. This implies that children with a medium to high self-concept cannot be distinguished from each other using the SPPC.

Furthermore, although each scale of the SPPC consists of 6 items, some constructs can be measured by 4 items (the PA scale), whereas other constructs need more items (the SC scale). There is, however, a catch. For the SPPC, it is clear that for the PA and GS scales there are only three items needed to obtain a scale with information around 12, but these items are semantically similar or have the same meaning for young children. Our example showed that Items 22 and 24 are semantically similar and that although Item 35 of the GS scale ("Happy with who I am") seems to have a different content, additional qualitative research showed that this item is interpreted by many children as "Happy with how I look like".

But scales should not be very short. First, when we use measurement precision as a criterion we select items that basically repeat the same question. Therefore, we distrust personality scales that consist of only a few items. Although we often encounter phrases like "given a standard error of 0.35, only two to three items are needed to evaluate a single dimension", we think that a researcher should be extremely careful when scaling persons on the basis of two or three items. Consider, for example, the subscale SA for boys. We found (roughly) the following item information values for $\hat{\theta}$ values in the range (-2, 0) for the six items (1, 3, 1, 0.4, 0.4, and 0.2). This implies that if a researcher accepts a standard error of 0.5, only one or two items are needed. However, good measurement is not only about precise measurement of the latent trait, it is also about replication and validity (in fact it is all about replication and validity).

What is often neglected is that very few items will result in an unacceptably low classification consistency (see e.g., Emons, Sijtsma, & Meijer, 2007). Classification consistency refers to the percentage of persons assigned to the same diagnostic category (such as low self-concept) by two hypothetical independent replications of the test. Emons, Sijtsma, & Meijer (2007) showed that for short tests (between 6 and 12 items) classification into two categories (e.g., treatment and nontreatment groups) resulted in at most 50% correct classifications, whereas results for longer tests (20 and 40 items) were much better. Although one may argue that short measures will never be used in practice to classify a person, for example, for a treatment, we noticed that especially in medicine and health psychology there is a trend to use extremely short scales, for example, through the Internet to classify persons as depressed or anxious (e.g., Taylor & Deane, 2002).

Thus, there is a dilemma when using inventories in personality measurement. On the one hand, we do not want to administer long self-reports that consist of items that basically repeat the same question over and over; on the other hand, inventories cannot be too short because results are difficult to replicate. What to do? A way out of this dilemma is perhaps the observation that in practice questionnaires are seldom used as the only indicator to classify a person into a category. Often interviews, observation, and other tests and questionnaires are being used. Combining information from different (sometimes unreliable) sources will result in acceptable replication rates and acceptable validity. Cronbach (1954) already showed that, given a fixed testing time, it is often better for personnel selection or treatment referral to use many short tests covering many dimensions than one (large) accurate test. This, thus, justifies the use of the SPPC as a general multidimensional measure of self-perception. At the same time, it warns the psychologist to overemphasize the

meaning of the total scores on the individual scales and thus - perhaps more important - subtest profiles, when they are not combined with other information. Furthermore, because researchers are often interested in measuring broad-band constructs it seems a good strategy, when constructing typical performance measures, to strive for a number of weak or medium Mokken scales. The alternative strategy to strive for strong Mokken scales consisting of many items will result in scales with items with similar content.

Because self-perception is a quasi-trait in a community sample, it has consequences for longitudinal research. For example, Shapka and Keating (2005) studied the longitudinal changes in multiple domains of self-concept over a 2-year period in adolescence. They found that most domains of self-concept increase with age, although perceived scholastic competence decreased. When SPPC scales are used in a longitudinal setting to measure change over time, it is very important to realize that the SPPC provides different measurement precision across the trait range. Clearly, the SPPC scales are a lot more sensitive to change in the lower regions of the latent trait because standard errors are much smaller there than in the higher regions of the scale.

A limitation of this study was the relatively small sample size for boys and girls for estimating the item parameters in the GRM analyses. Although we recognize this, we observe that the results of the GRM analysis were confirmed by the Mokken scale analysis for which less persons are needed. Also, our aim was not to provide definitive parameters estimates but rather to address specific questions regarding the strengths and weaknesses of the SPPC. Finally, in this study data were obtained from a nonclinical population. Because the SPPC is intended to be used in both clinical and nonclinical populations, future research may investigate the psychometric quality of the SPPC in a clinical population.

2.5 Recommendations

Given our perspective on the use of the SPPC in practice, we have the following recommendations and take-home messages:

(1) Psychologists and researchers should realize that contrary to earlier research, the subscales GS and PA are measuring similar concepts. Thus, the subscale GS should not be interpreted as an overarching scale, that is, as a scale that measures "general self-concept". Because of its high correlation with PA, it seems more of an appearance scale.

(2) Researchers should be very careful to use the total scores on the different subscales to distinguish children with a medium to high self-concept. Measurement in these ranges of the latent trait is very unreliable. Although, some subscales (PA) are suited to distinguish children with a low self-concept from children with a medium to high self-concept, in general the scales of the SPPC do not provide very accurate latent trait estimates. PA is a strong scale, whereas AC is a very weak scale. We, therefore, suggest that future research may reconsider a revision of several subscales of the SPPC. Using both content information as well as information from IRT analysis, items can be selected that allow for constructs that are broad enough to have any empirical validity and that provide acceptable measurement precision to distinguish children on the different important constructs of self-concept. We realize that it will not always be easy (perhaps sometimes impossible) to come up with subscales that are both reliable and that measure constructs that are broad enough for prediction, but we think that IRT can be of help in obtaining better SPPC scales than the existing ones.